THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 62, No. 7, September 1983
Printed in U.S.A.

On the Recognition of Isolated Digits From a Large Telephone Customer Population

By J. G. WILPON* and L. R. RABINER*

(Manuscript received February 25, 1983)

A field study was initiated to learn about the effects of various telephone transmission and switching conditions on the algorithms currently used in the Bell Laboratories, Linear Predictive Coding (LPC)-based, isolated word recognizer. Digit recordings were obtained from customers over a variety of transmission facilities. During a 23-day recording period a total of 11,035 isolated digits were recorded. For each recording, statistics were recorded about the line condition, the background environment, and the customer's ability to speak his/her telephone number as a sequence of isolated digits. Also recorded was information about the ability of the automatic word endpoint detector to find each spoken digit and to accurately determine the correct endpoints. The results of several recognition tests are presented—one using a previously defined set of laboratory-created digit reference templates, and several others using new sets of reference templates from a subset of the recorded digits. The performance of the recognizer is poor (average digit accuracy of 77.4 percent) using the laboratory template set, but improves substantially (average digit accuracy of 93.1 percent) for a template set created from the field recordings. The reasons for this improvement in digit recognition accuracy are presented, along with their implications to future work in isolated word recognition.

I. INTRODUCTION

Research on the problems involved in speech recognition has been carried out at Bell Laboratories for close to a decade.[1-7] In all these

*Bell Laboratories.

studies the speech database consisted of utterances recorded under laboratory conditions (i.e., cooperative subjects, soundproof booths, and subject prompting) using dialed-up lines over a local Private Branch Exchange (PBX). Peak signal-to-noise ratios ranged from 40 to 60 dB under these conditions. The recognition systems previously studied involved either a user training phase (speaker-dependent systems) or no training phase (speaker-independent systems). The vocabulary sizes ranged from as few as 10 words,[2] to as many as 1109 words.[7] Our past studies of speaker-independent systems involved a relatively small number of subjects, typically 100 for training and 10 to 40 for test calculations. Our current recognition systems performed very well given these conditions.[8]

To test the viability of speaker-independent, isolated word recognition systems for large user populations, it was necessary to conduct an experiment under "real world" conditions. Such an experiment involves using noncooperative telephone customers speaking in an uncontrolled environment over a set of randomly dialed telephone lines. This paper presents such an experiment and its implications of future speech recognition work. During the course of the experiment, a speech database was collected over a 60-day period in a Bell System environment (i.e., recordings obtained directly at a Bell System switching office) from over 3100 subjects. The vocabulary chosen for this study was the 10 digits, zero through nine (the digit zero was generally pronounced "oh"). Since we wanted a system that could handle a large number of users with the least amount of burden to the user we chose to make the system speaker independent.

There are several very important recognition issues that need to be resolved, and only by using a very large speech database, such as the one we obtained, can these issues be addressed. The most important issues involve training the recognizer. Since we have required that the recognition system be speaker independent, several questions arise as to how one obtains the necessary training data. Should training tokens be used from only the "best" speakers over the cleanest telephone lines, or should all the training samples be randomized, i.e., from any talker over any quality transmission line? Another issue involves the number of training tokens needed to adequately represent an extremely large number of potential users. In past studies we have used at most 100 representations for each vocabulary item. Do we necessarily need more tokens? In this paper we investigate these issues among others.

Our results indicate a distinct number of "real world" problems that must be considered when implementing a speech recognition system with a widespread applicability. These include properly treating highly variable background conditions, devising procedures for handling au-

tomatic endpoint detector failures, and problems associated with obtaining isolated speech input. Methods of handling these problems will also be discussed in this paper.

In Section II we describe how we obtained the speech recordings. Section III presents an evaluation of our current speech recognition system, on the basis of a series of recognition experiments. A discussion of the overall results and their implications is given in Section IV.

II. RECORDING PROCEDURE

Figure 1 depicts the overall recording setup used in this study. Recordings were made at a Bell System switching office in Portland, Maine, and then transmitted back to the Murray Hill Laboratory for analysis. The sequence of operations to record a single user's speech was as follows:

1. A site observer (SO) issued a prerecorded spoken message (a prompt) requesting that the user speak his or her telephone number as a sequence of isolated digits. After the first three digits were spoken,* the observer then initiated recording (i.e., digitizing the spoken data) of the spoken number sequence. As each of the digits was spoken, the observer entered it on a keyboard. Four digits were nominally recorded (digitally at a 6.67-kHz rate) for each subject.

2. The observer determined if the digit sequence was not spoken in an isolated format (i.e., spoken without sufficient pauses). If so, the observer initiated another prerecorded spoken message (a reprompt) requesting the user to repeat the number with a longer pause between digits. After the subject completed speaking the number, he or she was given a prerecorded "Thank you" message.

During this first phase an observer at the Murray Hill Laboratory (MHO) had also been monitoring the call and now took over handling the call, performing the following operations:

1. The MHO also determined (based on listening) whether the digit sequence was spoken in an isolated manner. If the MHO decided that the speech was unacceptable (either because it was spoken in a connected manner or because of unacceptably bad telephone line conditions), a reject code was entered and the entire procedure was terminated for the current call.

2. If the call was acceptable, the MHO entered the sequence of spoken digits heard. This sequence was compared with that entered by the SO and any discrepancies (errors) were noted and fixed by listening to the recorded digits.

---

* For reasons of customer privacy we were not allowed to record all seven digits of the telephone number.

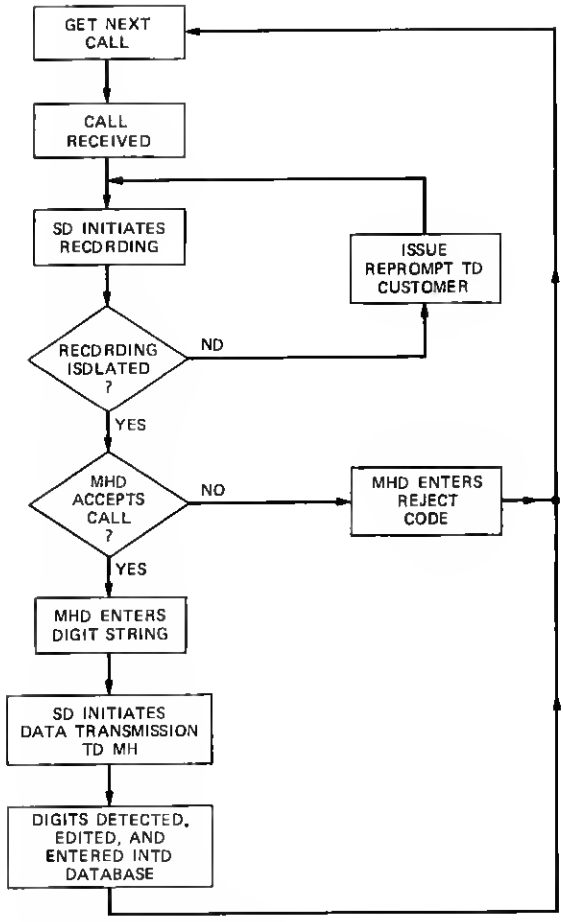

Fig. 1—The overall digit recording system.

3. At this point the MHO initiated digital transmission of the digitized speech from Portland to the MH laboratory.

Once the transfer was completed, an eighth-order Linear Predictive Coding (LPC) analysis, and automatic endpoint detection were performed.[9] The log energy of the waveform was displayed to the MHO, along with the automatically determined sets of endpoints indicating where in the recording interval the isolated words could be found. At this point the MHO had the option of modifying any or all sets of endpoints computed. The segmented speech was then entered into the database for later examination.

Using this procedure we recorded approximately 11,000 digits from

3100 subjects over a 23-day period. During the first 11 days no reprompting was used and recordings were taken for about 8 hours each day. Beginning wih the 12th day the reprompt procedure was instituted and we began recording for 12 hours each day. Figures 2 through 5 show the breakdown of the data recorded. Figure 2 shows a plot of number of digits recorded on a daily (session) basis, for males, females, and the total. For the first several days, only about 200 to 300 digits were recorded because both observers were learning. The dip in recordings around day 10 was due to a major snowstorm which curtailed recording. Figure 3 shows a histogram of the number of utterances recorded per digit for males, females, and total. The digits 2 and 3 bad the highest number of occurrences, and the digits 9 and 0 had the fewest number of occurrences. This phenomenon is due to the fact that the digits 0,9 (and sometimes 1) are reserved in some positions for pay phones, businesses, etc. The general falloff in number of occurrences from 2 to 8 is due to the manner in which the telephone numbers were assigned in the Portland area.

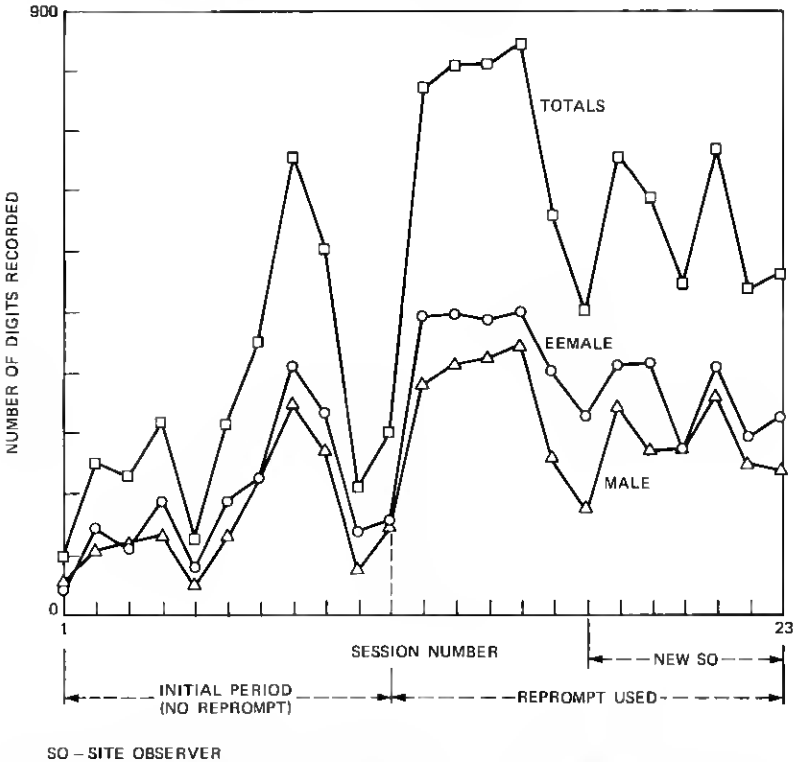

Fig. 2—The number of digits recorded as a function of session number for males, females, and combined (totals).

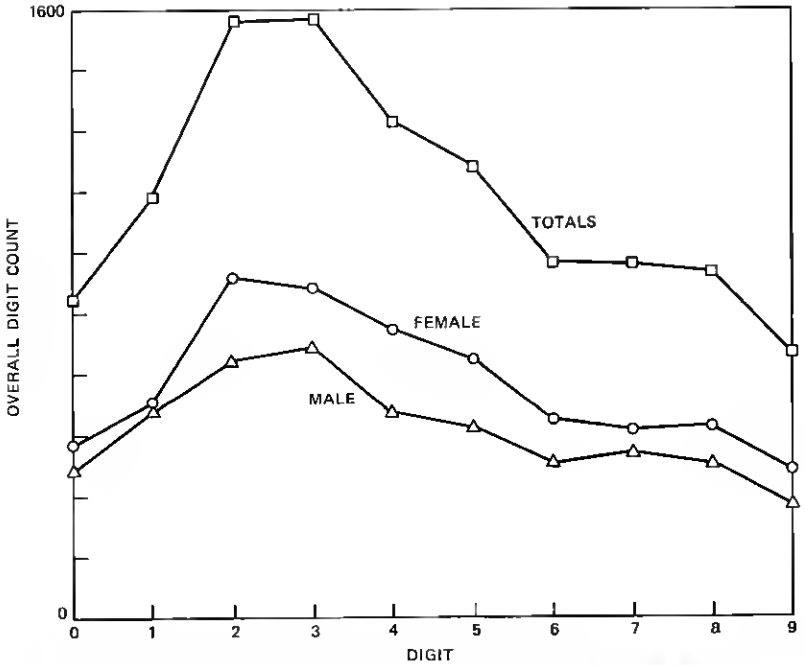

Fig. 3—Overall digit count as a function of the digit for males, females, and combined (totals).

There were several problems that occurred during the recording phase. These problems were classified as being in one of two groups. The first group contained problems associated with the telepbone transmission conditions. Artifacts such as clicks, tones, and hum were often superimposed on the subject's speech. Resulting peak signal-to-noise ratios varied from as little as 10 dB to as much as 60 dB. The second group consisted of problems related to the talker and the environment in which he or she spoke. These included nonisolation of speech (i.e., the digits were connected) and the presence of extraneous background speech, such as people talking in the background or a television set being played at an audible level at the handset. Most of the user failures were severe enough to warrant elimination of the customer's speech from the database.

Figure 4 shows a plot of the percentage of calls accepted (i.e., at least 1 digit was extracted from the call) as a function of the session number. The rejected calls are a sum of both SO and MHO rejections. We can see that only 53 percent of calls yielded at least 1 digit. There were several reasons for such a low yield. The main reason for rejection was nonisolation of speech. This took one of two forms; either the subjects spoke in pairs of digits (e.g., 43 followed by 27), or tbey connected all four digits (e.g., 4327). Figure 5 shows a plot of the

percentage of these two types of rejections, as a function of total rejections, on a per-session basis. During the initial phase of recording, digit pair rejections accounted for about 30 percent of all rejections. Similarly, fully connected speech accounted for approximately 27

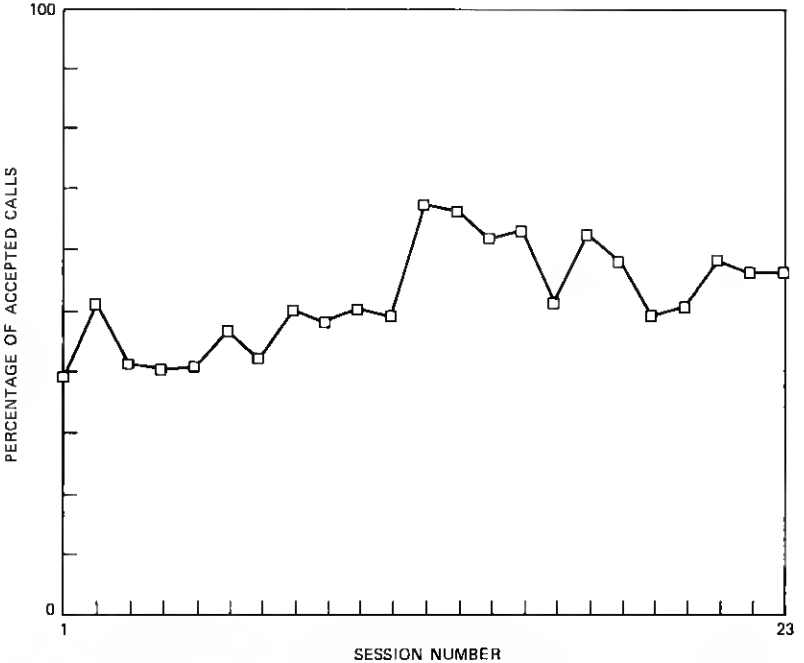

Fig. 4—The percentage of handled calls that were accepted as a function of session number.

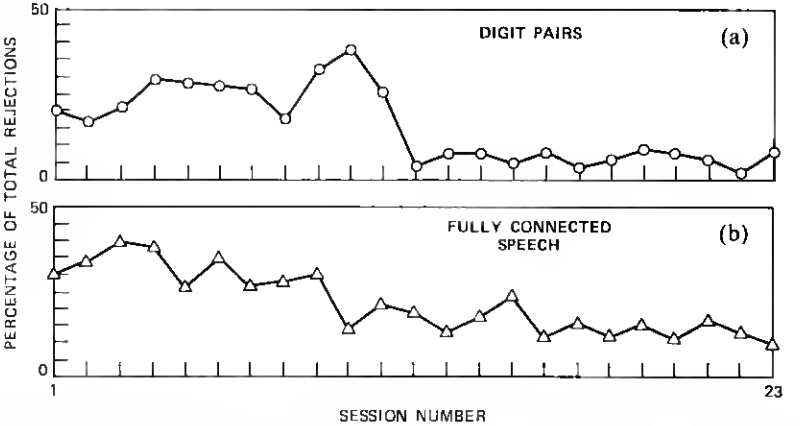

Fig. 5—Plots of percentage of total rejections because the number was spoken as (a) a set of digit pairs (b) as fully connected speech as a function of session number.

percent of total rejections. The plots show that a strong decrease in the number of these rejections occurred after the reprompting phase was initiated (Session 12), resulting in rejections rates of 9 and 15 percent for digit pairs and connected speech, respectively.

Another problem encountered during the recording process was the failure of the automatic endpoint detector to segment all the words properly. This algorithm[9] used the log energy contour of the speech and, after normalizing for hackground noise level, found as many digits as possihle in the recorded string. For "ideal" transmission conditions and correct speaking of the number (i.e., as a sequence of isolated digits) the task of digit detection is relatively straightforward and essentially makes no errors.[9] Figure 6a illustrates such a case for the digit string /5946/. This figure shows the log energy contour of the recording. The dashed vertical lines indicate heginning and ending frames for automatically detected digits. For this example the peak-signal-to-average-hackground-noise ratio was about 54 dB (i.e., the average signal-to-noise ratio was close to 40 dB). Furthermore, the background noise was fairly stationary and at a low level.

Unfortunately, most of the recordings differed substantially from that of Fig. 6a. Figures 6h through 6e illustrate some problems that made automatic reliahle detection of the digits very difficult. Figure 6h shows a log energy contour of the digit sequence /5282/ in which the signal level (during talking) was fairly low (peak-signal-to-hackground-noise ratio of 26 dB), and to further complicate matters, the background consisted of a mixture of noise and switching transients (clicks) generated within the telephone plant. For this sequence only three of the digits were automatically detected; the second digit was missed.

Figure 6c illustrates a case where the first three digits of the string /2383/ were spoken without a pause hetween the digits (i.e., in a connected format) and thus only one isolated digit could be used from this recording.

Figure 6d illustrates a case where the voice signal was corrupted hy continuous signaling tones throughout the recording interval. The tones were set at a sufficiently high level so that the peak signal-to-tone ratio was only ahout 35 dB. Although, for this example, the locations of the major portions of each of the four spoken digits (4672) were properly detected, the parameterization of the signal (used later to recognize the digits) was greatly distorted by the tones present while the digits were heing spoken. Furthermore, only the initial vowel portion of the digit six was located. The ending frication was lost in the tonal hackground.

Figure 6e shows an extreme case in which the background level of the line consisted of high-level noise and other extraneous sounds (i.e.,

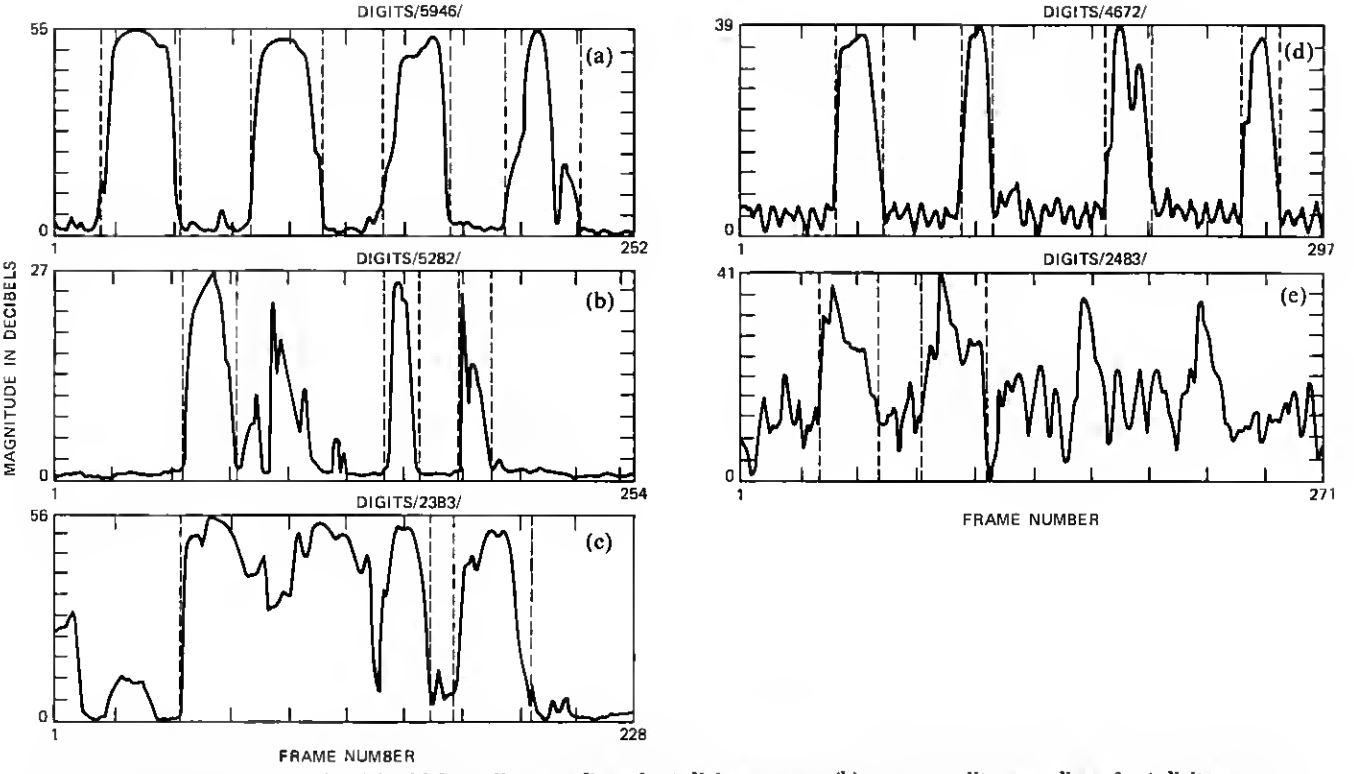

Fig. 6—Log energy contour for: (a) a high-quality recording of a 4-digit sequence; (b) a poor quality recording of a 4-digit sequence with low signal level and transmission clicks in the background; (c) a 4-digit sequence where the first three digits were not spoken in isolation; (d) a 4-digit sequence corrupted by continuous signaling tones throughout the recording interval; (e) a 4-digit sequence corrupted by high background noise and other extraneous sounds.

a very poor line). For this case it was impossible to detect accurate beginning and ending locations for any of the digits in the spoken string (/2483/).

The examples in Figs. 6d and 6e justify the importance of having the MHO check the accuracy of the automatic endpoint detector. In cases in which reliable endpoints are obtained automatically (no errors) the MHO allows the program to store the new digits in the database and update the relevant statistics. In all other cases the MHO can change endpoints or eliminate digits entirely. In this manner recordings with fewer than four isolated digits can provide one or more digits to the database and therefore the call is not entirely wasted. A discussion of the endpoint accuracy will be given in a later section.

An indication of how well the automatic endpoint detector performed was how often the MHO had to modify the endpoint sets. Figure 7 shows a plot of the percentage of calls requiring 0, 1, 2, 3, or 4 endpoint changes as a function of session number. This figure indicates that before using the reprompt, about 45 percent of the accepted calls needed no changes in word endpoints—i.e., no endpoint errors were made. After the reprompt was introduced the percentage rose to about 55 percent—i.e., a gain in endpoint accuracy of about 10 percent caused by the reprompt. This figure also shows that one set

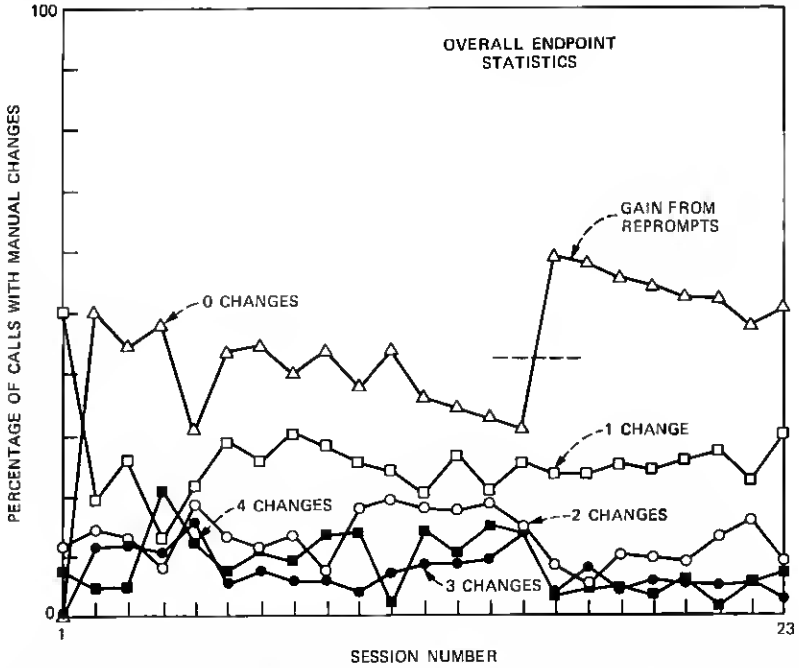

Fig. 7—Percentage of calls with from 0 to 4 sets of changed endpoints after manual corrections, as a function of session number.

of word endpoints needed to be modified about 25 percent of the time, and either three or four sets of endpoint modifications had to be made about 18 percent of the time. These results show clearly the necessity of improving the word endpoint detection. We are currently investigating into new methods of endpoint detection based on the types of problems encountered in this study.

III. ISOLATED DIGIT RECOGNITION EXPERIMENTS

3.1 Description of final database

After 23 days of recording over a two-month period, the final database contained 11,035 isolated digits taken from 3153 customers spoken during a wide range of telephone transmission conditions. Before any recognition experiments were performed, the full 11,035-digit database was listened to, and each digit was subjectively classified according to a set of background noise conditions to see if any one particular type of recording condition was harder for the recognition system to handle than another. Table I shows the individual types of conditions and the distribution of digits in the eight categories that were used. This table shows that 38.6 percent of the digits were classified as "acceptable". This meant that the background noise on the line was low (a signal-to-noise ratio in tbe range of 40 to 60 dB), and the given subject spoke in a clear, articulate voice, implying that the endpoint detector would have little problem with these utterances. These words were judged to be the "best" tokens, as close to laboratory data as was possible. The second category had similar signal-to-noise ratios, except here the callers did not speak in a normal fashion (presumably due to the novelty of being asked to speak in an isolated digit format). In such cases the customers dragged out words, or pronounced them in an abnormal way. This accounted for 10.3 percent of the digits. The remaining categories were used to describe the type of background noise present on the line. (If no background noise was present classes 1 and 2 were used.) About 12.5 percent of the digits came from strings that had a loud "crackling" noise superimposed on

Table I—Distribution of digits into categories based on background characteristics

Condition/Code	Count	Percent
Acceptable/1	4257	38.6
Customer-Related Problems/2	1133	10.3
Crackling Noise/3	1379	12.5
Pops, Clicks/4	1741	15.8
Tones, Whines/5	1386	12.6
Hum/6	752	6.8
Whirring/7	129	1.2
Background Speech/8	258	2.3
Totals	11035	100

the speech. Another 15.8 percent had pops and/or clicks throughout the recording. About 12.6 percent of the digits had loud tones present, mostly at 2600 Hz; and another 6.8 percent had loud "humming" noises superimposed with the speech from 200 to 400 Hz. We classified about 1.2 percent of digits as having a noise between "crackling" and "bum". (After we listened further we realized that this category could have been eliminated and its members classified as "crackling" noise.) The final 2.3 percent of the digits had background speech present—either other people's conversations, or television or radio sounds.

Table II shows the distribution of the 3153 calls, categorized by the number of digits (1, 2, 3, or 4) that were obtained from the calls. We see that for 68.8 percent of the calls four digits were actually obtained from the subject's spoken input. The average string length was 3.5 digits.

3.2 *Description of recognition experiments*

To determine how well we could recognize digits from this database, eight isolated word speech recognition experiments were performed. In the first experiment, the template set consisted of a set of 12 speaker-independent templates for each of the 10 digits plus the word "oh" (since the majority of talkers used "oh" instead of "zero" for the digit 0). The template set (MH templates) was obtained several years earlier from a clustering analysis of the speech of 100 talkers (50 male, 50 female) with recordings having been made over a local, dialed-up telephone line, providing from 40 to 60 dB peak-signal-to-noise ratios for all recordings.[2] The recognizer was the LPC-based recognizer, which has been in use in the Acoustics Research Department for several years.[1,2]

Figure 8 shows the results of this first recognition experiment. Figure 8a shows plots of the recognition accuracy for the top word candidate and the top two-word candidates as a function of the session number. The overall average word accuracy for the top candidate was 77.4 percent and was basically steady (to within statistical variations) with time.

Figure 8b shows a plot of the overall individual variations in digit accuracy. The digit 8 attained an accuracy of close to 95 percent, whereas the digit 4 was only 50 percent accurate. The major problem

Table II—Distribution of calls yielding 1, 2, 3, or 4 digits

	No. of Digits in String*				
	1	2	3	4	Total
Count	109	375	501	2168	3153
Percent	3.4	11.9	15.9	68.8	100

* Average string length −3.5 digits.

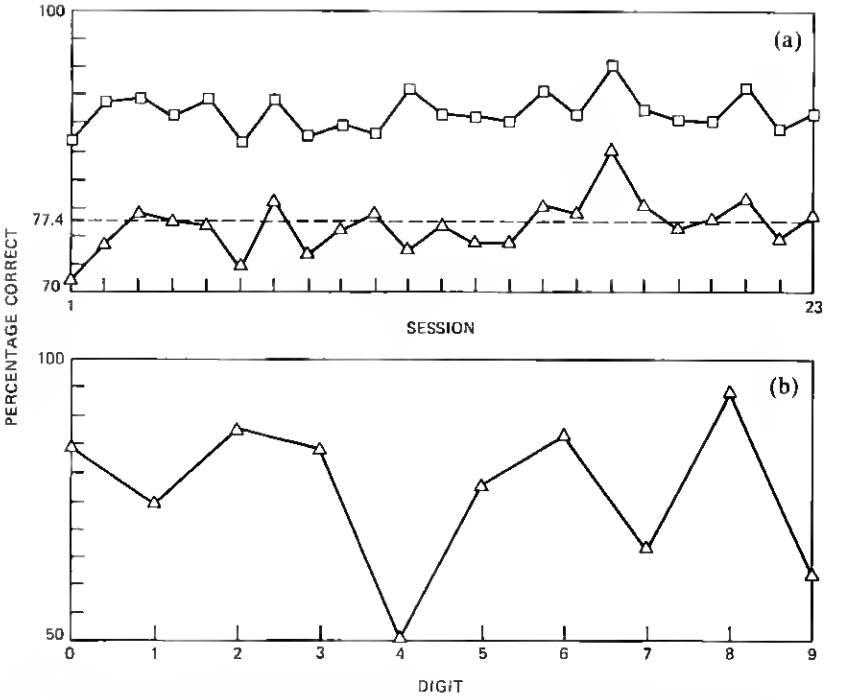

Fig. 8—Average digit recognition accuracy from: (a) MH templates for the top candidate and the top two candidates as a function of session number; (b) MH templates as a function of the spoken digit.

with the digit 4 was that about half the talkers pronounced it as /foe/, rather than /fore/, as represented in the template set. Hence for all such cases the digit 4 was recognized as 0 (i.e., the templates for "oh" provided the best match). Similarly, a modest number of confusions were found between the digits 5 and 9, and 1 and 9.

There was a basic problem with the training set of the first experiment. The pronunciations of the digits, as were prevalent in the Portland, Maine, area, were not well represented within the templates, and the difficult noise recording conditions led to large analysis degradations. Therefore, a second recognition experiment was run in which about 35 percent of the database was selected at random on a per-digit basis as a training set from which a new set of word reference templates [Portland (PO) random templates] was created. Reference template sets were generated using the UWA clustering algorithm of Rabiner and Wilpon,[6] yielding 12-, 20-, 25-, and 30-template-per-word sets, with the 30-template-per-word set yielding the best recognition results. The entire database of 11,035 digits was again used as a test set and the recognition results are given in Figs. 9 through 11 for the 30-template-per-word set. Figure 9a shows plots of the average word

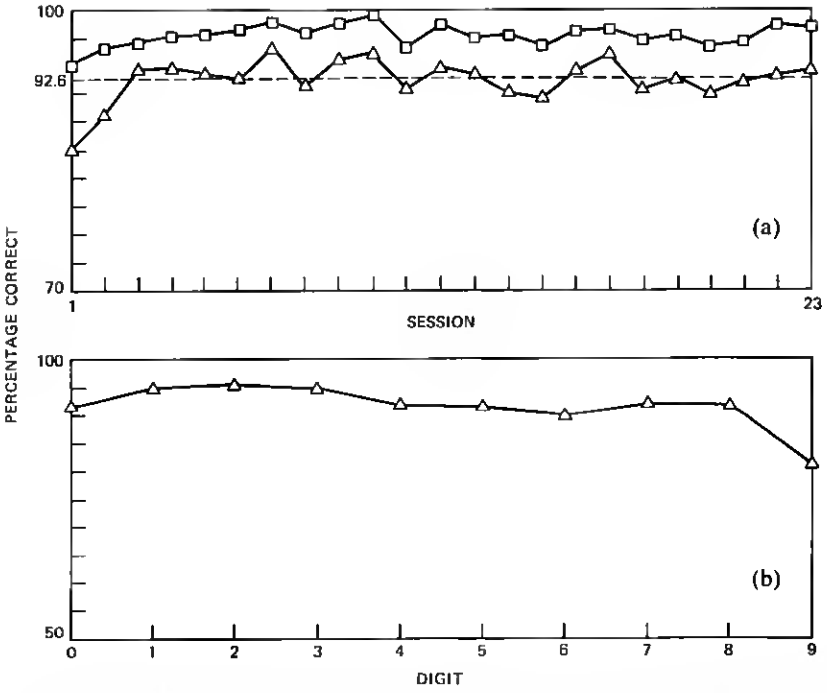

Fig. 9—Average digit recognition accuracy from: (a) PO random templates for the top candidate and the top two candidates as a function of session number; (b) PO random templates as a function of the spoken digit. (The same scale is used in Fig. 8b for comparison.)

recognition accuracy as a function of session number for both the top and the top two word candidates. The overall average word recognition accuracy for the top candidate was 92.6 percent and the individual session scores were fairly constant in time.

Figure 9b shows the overall recognition rates for the individual digits. We now see that all digits, except 9, were recognized with greater than 90 percent accuracy. Interestingly, the digit 9 was the one with the fewest tokens in the training set; hence improved recognition on 9 might result from a larger training set.

Figure 10 shows a plot of the overall word recognition accuracy as a function of the number of templates used per digit. The recognition rate is essentially flat for about 18 or more templates per word; hence only small reductions in accuracy would result from reducing the computation by almost one half.

Finally, Fig. 11 shows the individual accuracy scores for each digit as a function of the number of templates per digit. For some digits, e.g., 4, 6, 8, the recognition accuracy saturates for a small number of templates per digit, whereas for other digits, e.g., 0, 5, 7, 8, a large

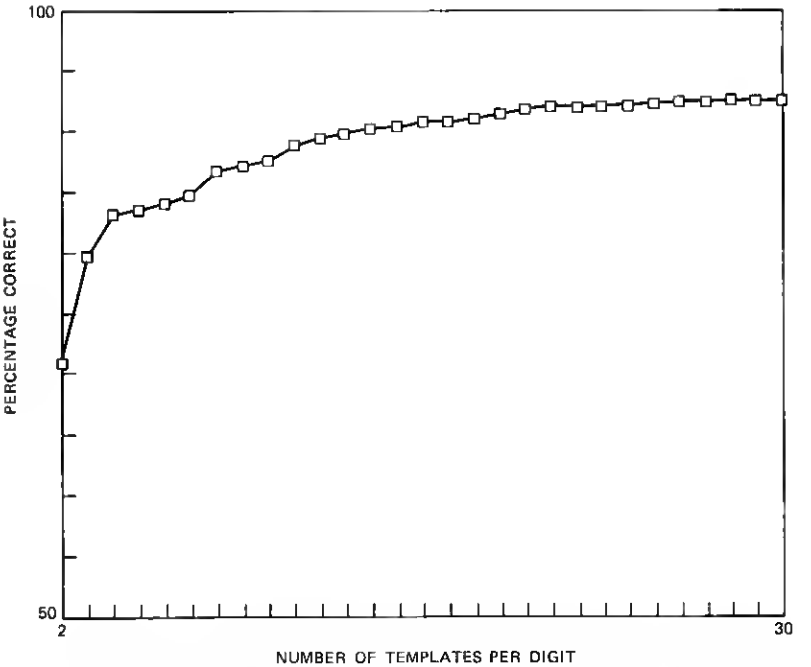

Fig. 10—Average digit recognition accuracy from PO random templates as a function of the number of templates used per digit.

number of templates per digit are needed. These results suggest that a reference set with a variable number of templates per digit could conceivably perform as well as the 30-template-per-digit set.

A problem in using a random set of utterances to train the system is the wide variability of transmission conditions present in the training tokens. The clustering procedure used to generate the templates tries to split the different spoken versions of a word into "similar" groups. If we now add an independent component to the speech, namely transmission variability, one would expect clusters also to be formed based on similarities in background conditions. Therefore we proposed the following recognition test. We used all the data that had been classified as "acceptable" (about 35 percent of the entire database) as a training set, from which we obtained a set of word reference tokens (PO clean templates). Since we were interested in a comparison with the PO random template experiment, a total of 30 reference templates were created for each digit. Again, the entire 11,035-word database was used as a testing set. The recognition results are summarized in Figs. 12 and 13. Figure 12a shows plots of recognition accuracy as a function of session number for the top and the top two word candidates. The average recognition accuracy over all days was

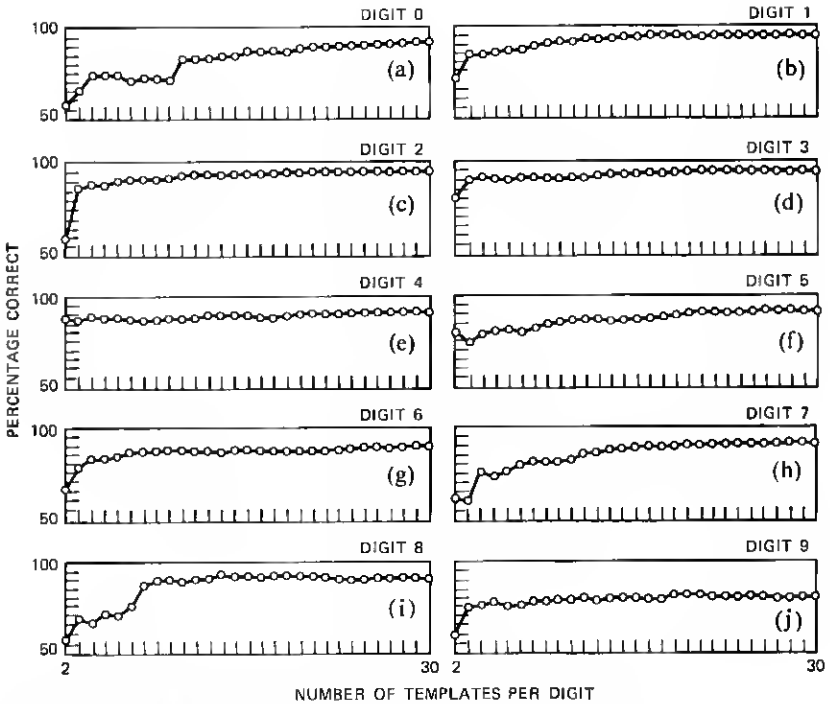

Fig. 11—Average recognition accuracy from PO random templates, for each individual digit, as a function of the number of templates used per digit.

93.1 percent. Compared with the PO random templates the difference in recognition accuracy was only 0.6 percent. Several inferences can be made from this result. First, it may not be necessary to go through the very time-consuming job of listening to several thousand spoken digit sequences and extracting only the "best" for use in training. Second, this result may indicate that the population of approximately 3900 training tokens was large enough to encompass all speaker variations (for most digits) and also all types of transmission system degradations.

Figure 12b shows a plot of the individual digit recognition scores. We see that the digit 9 still has a lower accuracy than the other digits. Figure 13a shows digit recognition accuracy as a function of condition code for tbe condition codes described previously. An accuracy of 96.6 percent was obtained when the testing set was also from the acceptable category (i.e., the same as the training data). The worst recognition results (86.7 percent) came from the digits categorized as having a very pronounced "humming" sound present during recording. Figure 13b shows string recognition accuracy as a function of condition code. The highest accuracy was 89.1 percent for "acceptable" data and the

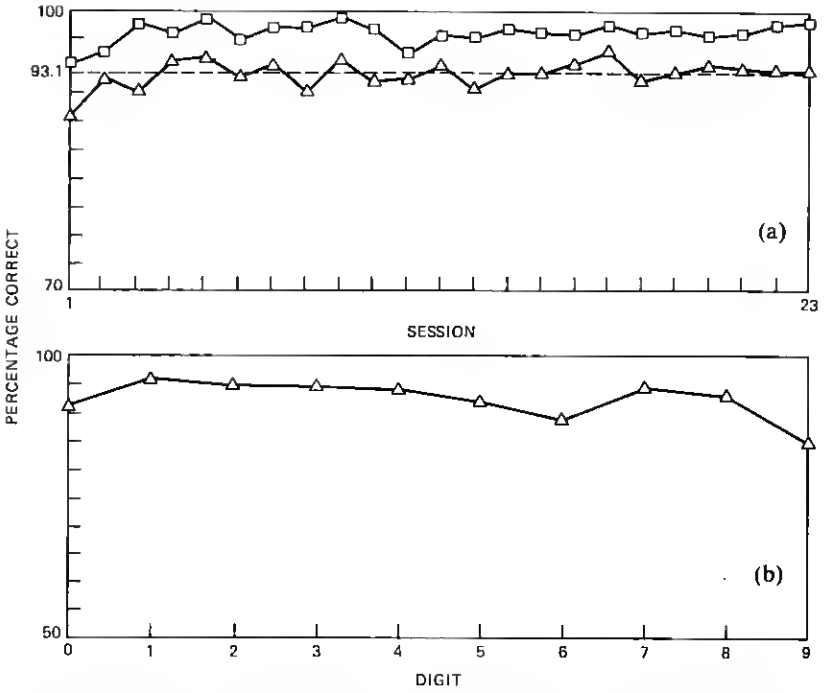

Fig. 12—Average digit recognition accuracy from the (a) PO clean template set for the top and top two candidates as a function of session number; (b) as a function of the spoken digit.

lowest accuracy was 64.3 percent for strings within the "hum" classification. The average string accuracy over all conditions was 80.8 percent. It was clear from these figures that the recognition errors were not distributed uniformly over all recording conditions. As a result, we were interested in determining whether digit errors were independent within each classification. Dependency might indicate that some sort of syntax could be applied to improve overall string accuracy.

To test the assumption that digit errors were independent within each string and classification category, a simple Bernoulli model was assumed in which the probability of error of a single digit was called α. Under these conditions, the probability, P, of four correct recognitions within a string of four digits is

$$\begin{aligned} P(\text{4 correct}) &= 1 - P(\text{single error}) + P(\text{double error}) \\ &\quad - P(\text{triple error}) + P(\text{quadruple error}) \\ &= 1 - 4\alpha + 6\alpha^2 - 4\alpha^3 + \alpha^4. \end{aligned} \tag{1}$$

For small α's eq. (1) can be approximated as

$$P(4 \text{ correct}) \approx 1 - 4\alpha. \tag{2}$$

Given that the average string length over all strings recorded was 3.5 digits

$$P(\text{string correct}) \approx 1 - 3.5\alpha. \tag{3}$$

In testing this hypothesis if we look at Fig. 13a, and multiply the digit error rates by 3.5 for each condition, we immediately see that eq. (3) above does indeed hold (at least to a reasonable level of accuracy). This indicates that within each classification the errors were probably distributed uniformly and randomly.

The next series of recognition experiments was carried out to determine whether we were using too many training tokens and therefore could reduce the amount of training data collected in future studies. Reference sets were created using one-third, one-half, and two-thirds of the data used in creating the 30-template-per-word PO clean template set (about 11.5, 17.2, and 23 percent of the entire database, respectively). The results are given in Tables III and IV. Table III shows the number of tokens in each training set or each digit, and Table IV shows the per-digit minimum, maximum, and

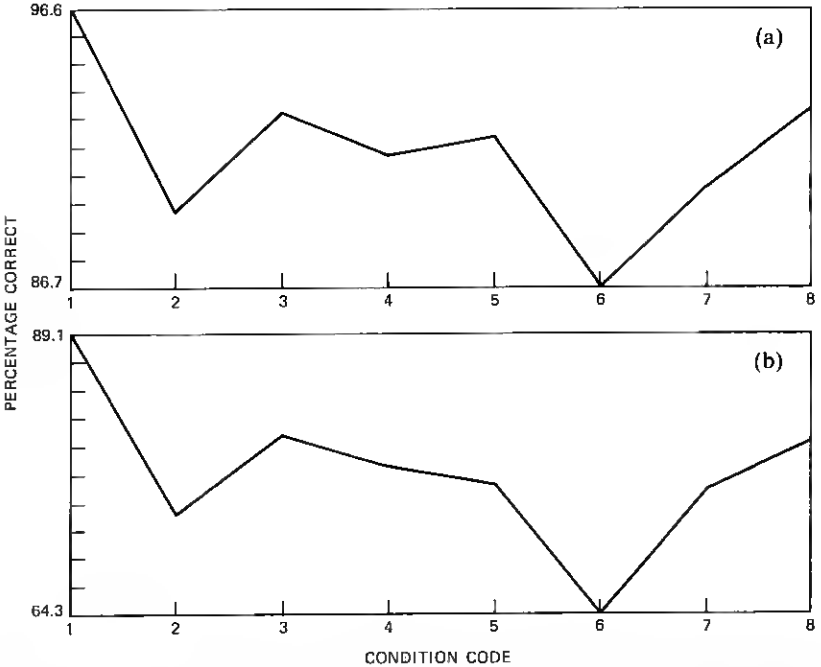

Fig. 13—Recognition accuracy as a function of PO clean template set for the: (a) acoustic classification on a per-digit basis; (b) acoustic classification on a string basis.

average accuracies for each experiment performed. The average per-digit recognition accuracy for the 12-template-per-word PO clean (one-third) template set was 89.7 percent for the 20-template-per-word PO clean (one-half) template set it was 91.4 percent; and for the 25-template-per-word PO clean (two-thirds) template set it was 92.2 percent.

The results show that the digit 9 consistently had the worst recognition scores. One possible explanation for this could be that in all cases the digit 9 had the smallest training set. The next recognition experiment was run to test this hypothesis. In this experiment we used a fixed number of training tokens per word. We were interested in seeing whether the 9 scores improved relative to the other digits. (Obviously we did not expect the overall accuracy to improve.) The results of this experiment are given in Table IV. We can see that the recognition accuracy for 9 has not changed (compared to the PO clean

Table III—Distribution of training tokens for individual words for each recognition experiment

	MH	PO Random	PO Clean	PO Clean (1/3)	PO Clean (1/2)	PO Clean (2/3)	PO Clean (Fixed)	PO Clean Variable
0	100	306	280	93	140	186	233	280
1	100	450	391	130	195	260	233	391
2	100	576	519	173	260	346	233	519
3	100	540	562	187	281	374	233	562
4	100	432	462	154	231	208	233	462
5	100	360	415	138	208	272	233	415
6	100	306	301	100	150	200	233	301
7	100	360	325	108	162	216	233	325
8	100	306	314	104	157	208	233	314
9	100	234	233	77	166	154	233	233

Table IV—Summary of recognition accuracies for the eight recognition experiments

	MH	PO Random	PO Clean	PO Clean (1/3)	PO Clean (1/2)	PO Clean (2/3)	PO Clean Fixed	PO Clean Variable
0	84.6	91.5	91.1	85.9	89.3	90.0	92.5	93.3
1	75.0	94.9	96.0	90.0	94.7	95.0	95.5	94.8
2	87.8	95.8	94.9	95.0	94.5	95.5	94.4	93.1
3	84.3	95.2	94.7	94.5	95.2	95.5	94.6	95.0
4	50.5	92.3	94.3	91.6	91.1	90.8	89.7	93.2
5	78.0	91.9	92.2	86.8	89.3	90.8	90.2	91.5
6	86.8	90.3	89.1	93.9	90.0	89.0	89.0	88.8
7	66.9	92.4	94.7	86.0	91.3	92.7	92.3	95.3
8	94.6	92.1	93.6	92.1	88.7	94.7	93.1	93.5
9	62.3	81.6	85.3	82.3	83.3	80.9	85.6	84.7
Average	77.4	92.6	93.1	89.7	91.4	92.2	92.1	92.6
Minimum (Word)	50.5(4)	81.6(9)	85.3(9)	82.3(9)	83.3(9)	80.9(9)	85.6(9)	84.7(9)
Maximum (Word)	94.6(8)	95.8(2)	96.0(1)	95.0(2)	95.2(3)	95.5(2)	95.5(1)	95.3(7)

template set), while the accuracies for the other digits have fallen. These results imply that some digits, e.g., 2, 3, need significantly fewer training tokens then do other digits, like 9.

In the final recognition experiment, we investigated the use of a variable number of templates per word. Tbe number of templates per word was cbosen on the hasis of the curves of Fig. 11 at the point where tbe recognition accuracy leveled off. The training set for this experiment was the same as used in the PO clean template set, namely, approximately 35 percent of the entire database for eacb of the 10 digits. The average recognition accuracy using this new template set was 92.6 percent. This was only 0.5 percent worse than the recognition accuracy of the PO clean template set, but with 100 fewer templates (i.e., about two-thirds of the computation).

3.3 Imposing a rejection threshold

Figures 14 through 16 demonstrate the effects of imposing a threshold on the recognition task. A recognition distance score ahove the

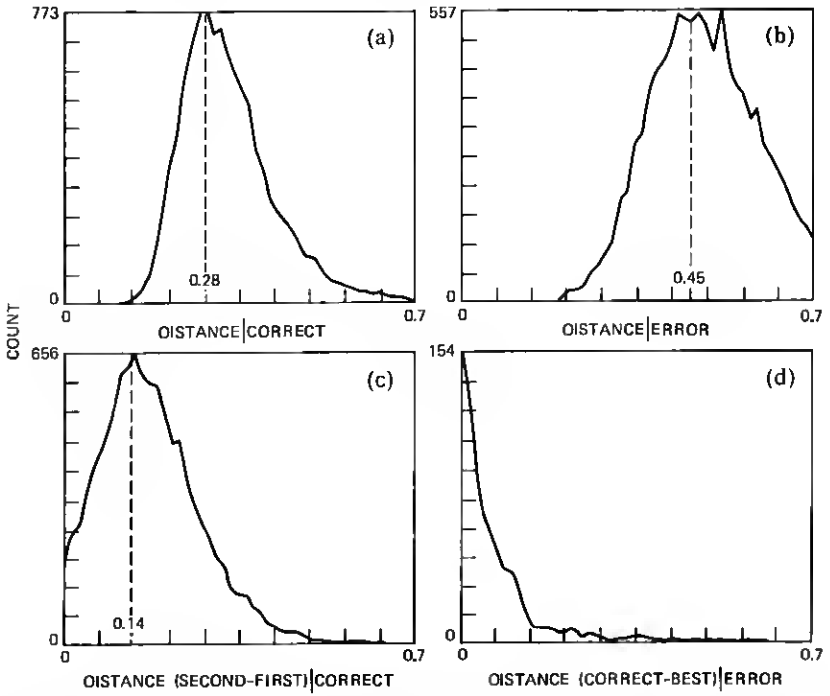

Fig. 14—Plot of histograms of LPC distances for (a) tbe correct word; (b) tbe closest incorrect word; (c) the difference between the best and second best cboice, given the best choice is correct; (d) the difference between the correct word and the top recognition candidate, given the top candidate is not tbe correct word.

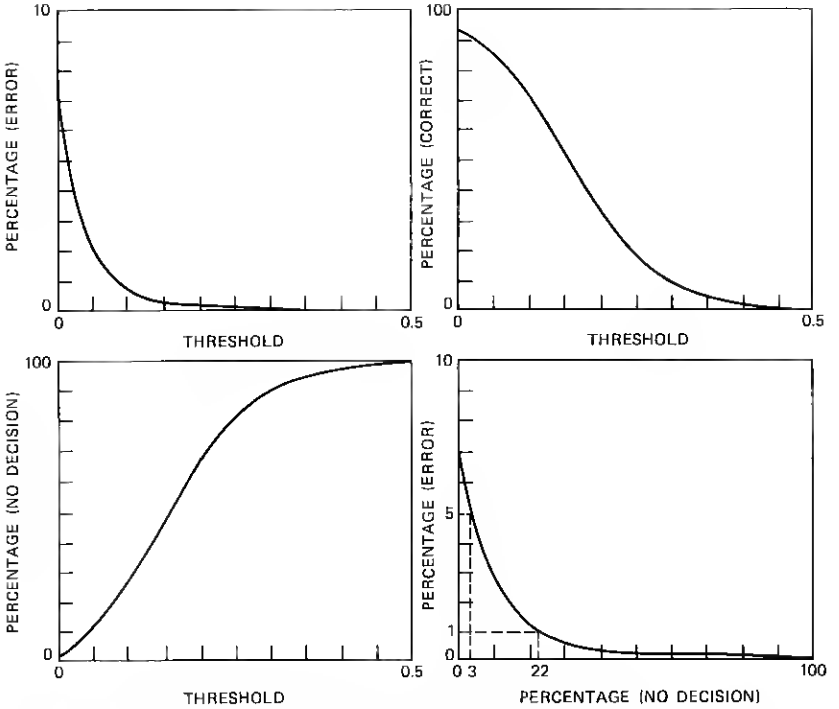

Fig. 15—Plot showing trade-off between no-decision rate and error rate on a per-digit basis, given a thresholding scheme imposed on the recognizer.

threshold would lead to a result of no decision by the recognizer. Figure 14a shows a histogram of LPC distances for the correct word (i.e., all 11,035 digits), with a mean correct distance of 0.28. Figure 14b shows a histogram of the scores for the closest incorrect word, with a mean distance of 0.45. Figure 14c shows a histogram of scores of the difference between the best choice and the second best choice, given the best choice is correct. Its mean of 0.14 indicates that when a word is recognized correctly the next closest word will on average have a distance score about 50 percent greater. Since LPC distances are on a log scale, this difference is a relatively large one. Figure 14d shows a histogram of the difference between the correct word and the top recognition candidate given the top candidate is not the correct word. This plot shows that when a word is misrecognized, the correct word has an LPC distance very close to that of the best choice. These results imply that a thresholding scheme can be applied to the recognition system and yield a good trade-off between accuracy and no-decision choices. Figure 15 shows the results of implementing such a thresholding technique. This plot shows the percent no decisions versus percent error rate, for the PO clean template set. We see that

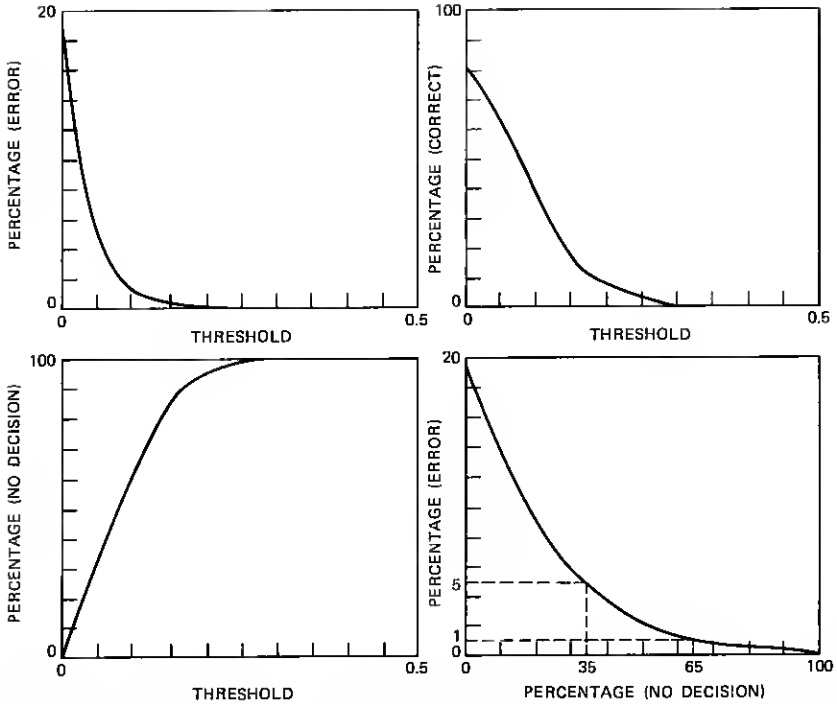

Fig. 16—Plot showing trade-off between no-decision rate and error rate on a string basis, given a thresholding scheme imposed on the recognizer.

if the task for which this recognition system is to be used can only tolerate a 1-percent per-digit error rate, a no-decision rate of 22 percent must also be accepted. However, a 5-percent probability of error yielded only 3 percent no decisions. Figure 16 shows a similar set of trade-off curves for full digit strings. For a digit string with an average of 3.5 digits, a 1-percent string error rate leads to a 65-percent no-decision rate and a 5-percent string error rate leads to a 35-percent no-decison rate. These results suggest that even though thresholding can be used to reduce error rate, other methods that do not lend themselves to such high no-decision rates should be investigated.

IV. DISCUSSION

The results presented in Sections III show that:

1. Significant problems exist in getting casual telephone customers to speak telephone numbers as a series of isolated digits. These are related to human factors issues (people don't normally speak in isolated word sequences) and problems induced by a wide variety of transmission and switching conditions.

2. The ability to detect words automatically in noisy or nonideal

environments is not adequate for about 50 percent of the recordings. There are some obvious ways of improving the current algorithm for finding words; and the database we collected in this test is currently being used to test various modifications of the algorithm.

3. The recognition results given in Table IV show that the accuracy of the recognizer, even when trained on a subset of the test data, is at best marginally acceptable with a maximum accuracy of 93.1 percent. However, some digits had recognition accuracies well over 93.1 percent, while others (i.e., the digit 9) had scores around 85 percent. More detailed analyses of the type of errors and their causes must be undertaken and some improvements in the training and recognition procedures must be made to approach the accuracies obtained for laboratory recordings (close to 98 percent).

Each of the above problems will be investigated carefully to improve all aspects of automatic recording, detection, and recognition of isolated words.

Another data collection exercise will begin soon at another site. Such issues as the effects of regional dialect on the reference templates will be studied. In addition, further testing of an improved endpoint detector will be performed.

V. SUMMARY

In this paper we have presented recognition results obtained from a speech database, consisting of 11,035 isolated digits, collected in an actual telephone environment from 3100 nonsolicited subjects. Several different reference template sets were used with a maximum recognition accuracy reported at 93.1 percent.

VI. ACKNOWLEDGMENTS

The authors would like to acknowledge the many individuals involved in collecting the data described in this paper. In particular, Eddie Youngs, Cindy Karhan, Marty Viets and Helen Hollinka provided the interface to the Bell System and were responsible for direct recording of the spoken digit sequences. Frank Pirz and Keith Bauer designed, built, and programmed tbe special-purpose recording hardware. Finally, Don Bock and Sandy MacNeil aided in the collection, processing, and analysis of tbe customer data at the Murray Hill laboratory. To each of these individuals we express our thanks for their help.

REFERENCES

1. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-23*, No. 1 (February 1975), pp. 67–72.
2. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker

Independent Recognition of Isolated Words Using Clustering Techniques," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-27*, No. 4 (August 1979), pp. 336–49.
3. L. R. Rabiner and J. G. Wilpon, "Speaker Independent, Isolated Word Recognition for a Moderate Size (54 word) Vocabulary," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-27*, No. 6 (December 1979), pp. 583–7.
4. J. G. Wilpon, L. R. Rabiner, and A. F. Bergb, "Speaker Independent Isolated Word Recognition Using a 129-Word Airline Vocabulary," J. Acoust. Soc. Amer., *72*, No. 2 (August 1982), pp. 390–6.
5. L. R. Rabiner and J. G. Wilpon, "A Simplified, Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," J. Acoust. Soc. Amer., *68*, No. 5 (October 1980), pp. 1069–70.
6. L. R. Rabiner and J. G. Wilpon, "Considerations in Applying Clustering Techniques to Speaker Independent Word Recognition," J. Acoust. Soc. Amer., *66*, No. 3 (September 1979), pp. 663–73.
7. L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and W. J. Keilin, "Isolated Word Recognition for Large Vocabularies," B.S.T.J., *61*, No. 10, Part 1, (December 1982), pp. 2989–3005.
8. L. R. Rabiner and S. E. Levinson, "Isolated and Connected Recognition—Theory and Selected Applications," IEEE Trans. Commun., *29*, No. 5 (May 1981), pp. 621–59.
9. L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-29*, (August 1981), pp. 777–85.

AUTHORS

Jay G. Wilpon, B.S., A.B. (cum laude) in Mathematics and Economics, respectively, 1977, Lafayette College, Easton, Pa; M.S. (Electrical Engineering/Computer Science), 1982, Stevens Institute of Technology, Hoboken, N.J.; Bell Laboratories, 1977—. Since June 1977 Mr. Wilpon has been with the Acoustics Research Department at Bell Laboratories, Murray Hill, N.J., where he is a Member of the Technical Staff. He has been engaged in speech communications research and is presently concentrating on problems of speech recognition.

Lawrence R. Rabiner, S.B. and S.M., 1964, Ph.D., 1967 (Electrical Engineering), The Massachusetts Institute of Technology; Bell Laboratories, 1962—. From 1962 through 1964, Mr. Rabiner participated in the cooperative plan in electrical engineering at Bell Laboratories. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently, he is engaged in research on speech communications and digital signal processing techniques. He is coautbor of *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Prentice-Hall, 1978), and *Multirate Digital Signal Processing* (Prentice-Hall, 1983). Former President, IEEE, ASSP Society; former Associate Editor, ASSP Transactions; former member, Technical Committee on Speech Communication of the Acoustical Society, ASSP Technical Committee on Speech Communication; Member, IEEE Proceedings Editorial Board, Eta Kappa Nu, Sigma Xi, Tau Beta Pi. Fellow, Acoustical Society of America, IEEE.